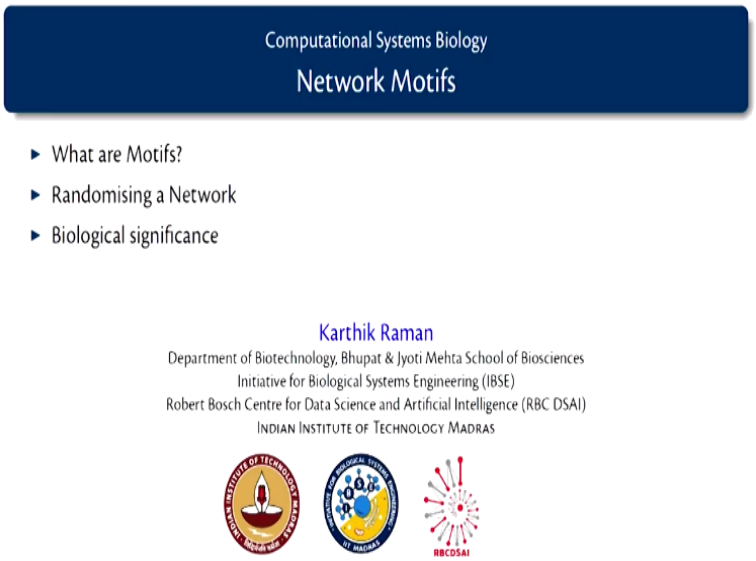
Computational Systems Biology
Karthik Raman
Department of Biotechnology
Indian Institute of Technology – Madras

Lecture - 24
Network Motifs

(Refer Slide Time: 00:12)



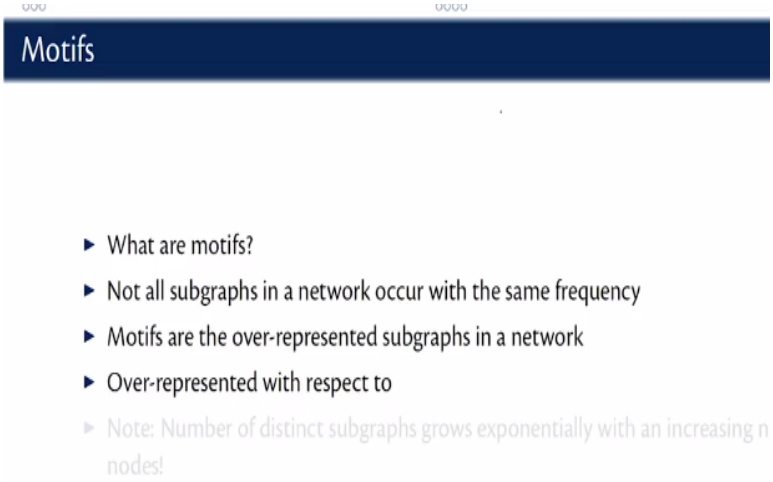
In today's video, we will look at network Motifs. I will introduce you to the concept of a Motifs in networks which we have already seen a little while ago, but we will focus a little more on how do you identify Motifs from a network and that involves identifying a null model, randomization and so on. What are the different types of methods that are used to randomize a network and what is the biological significance of these Motifs?

(Refer Slide Time: 00:42)

MOTIFS

So now let us look at Motifs. So what is a Motif? We said it is some sort of a pattern, but we already defined it as an over represented subgraph or rather a statistically significantly over represented subgraph in a network.

(Refer Slide Time: 00:54)



So not all subgraphs occur with the same frequency in any given network. You will find that some motifs are over represented, but over-represented with respect to what? With respect to some null model. So what becomes a null model? So you can have different null models. You could think of a Barabási-Albert null model or you could think up of different null models so it dependence upon what is the inference you want to draw.

This is where I need you to recall the statistics, because this is the very statistical kind of argument. So the common example I often gave is when you are testing out a drug. Let us say you want to find out a paracetamol works against headache. How do you test if paracetamol works against headache? **"Professor - student conversation starts"** You first establish a group of patients. You probably want to actually establish a group that is very homogeneous.

You know it has got to have like 10 boys and 10 girls, may be all between the age of 20 and 25, or whatever. It depends upon what is that you want it is, but let us you want to keep the things, the unknowns as narrow as possible. You do not want to have everybody from the age of 10 to 60 or something like that. You want to keep a narrow control group and you want to have a narrow group which you are going to test the drug on and how do you test it what would you give the control group.

A Placebo. So what is a placebo? Something that is not the drug, but does it have to have anything else. It must look like a drug. That is not important, but what is important is it should have everything in the tablet other than paracetamol. **"Professor - student conversation ends"** So we had some colouring agent, some other excipient that holds the tablet together. Whatever else is there in the drug may be as a colour so even that colour should be there because then you cannot say that this colour was not curing headache.

See finally paracetamol and the colour are there in your drug whereas paracetamol and colour were there in your control you cannot still decouple the effect of paracetamol from the effect of the drug. Why I am giving you this example is you can only answer the question based on the control that you are set up. After this you have find that 9 out of 10 people who were given paracetamol got cured of headache and only 2 out of 10 people who had the placebo got cured of headache.

You can now say that some confidence you have to a statistical test that those who took paracetamol, paracetamol actually works against headache. So in the same mean here you have to set up the right control. What does it that you want to ask of this network. So do you want to

look at networks that have the same number of nodes? So you could then say that in a node with in a network with 1000 of nodes and very likely to find this triangles subgraph.

Or the straight forward loop that I showed you yesterday, but is that what you want to do. Do you want keep something else constant? You may want to keep the degree distribution constant or you may want to keep the function constant if you are looking at a biological network because then you can answer suppose you only keep the degree distribution constant and allow for the function to change.

It is not easy to measure the function of a network, but let us assume that you have a way to say what is the function of your network whether it could be a metabolic network, then it is easy to say whether this network will grow on a particular nutrient source or not. So then if you allow the network function to not change you can now say that if you allow the network function to change you can say that these Motifs will be present in any functioning biological network.

These motifs are necessary for function to happen in the first place, but if you establish the appropriate control if you allow for function to not change and still your motifs is enriched then you can say that this motifs seems to be very unique to my network to the network that I am studying. It does not have to be a feature of all biological networks or all functional biological networks, but it seems to be a very special property of my network.

This is usually something that you want to fish out of a data set or out of any given network. So what is the statistical power that you have? So it depends upon the null model you establish or you know the null hypothesis, alternate hypothesis, good hold statistical testing. You should try and brush up on your basic statistics. So in this case what we want to study is how does my motif vary with respect to some null model? How do I now construct those null models?

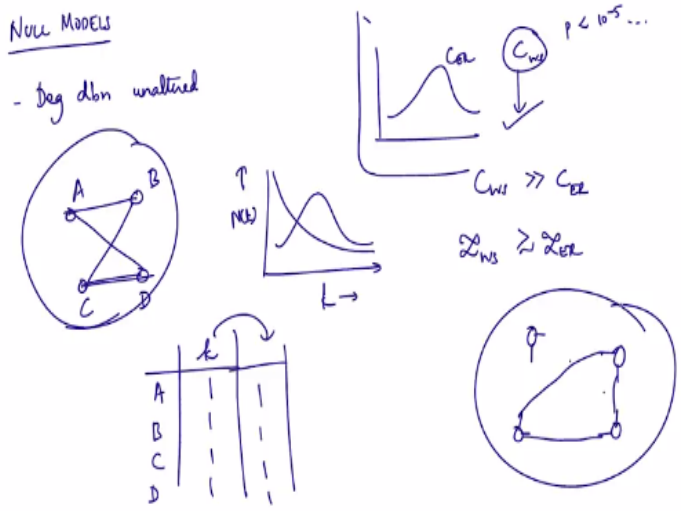
It depends. It depends upon the question. So as always in systems biology the modeling choice or the algorithm choice always depends. There is no 1 answer to what is it that you will use. It depends upon the question you want to answer. Do I want to identify what motifs are going to be present in this network in biological networks or in my particular network? Depending upon the

resolution of your question the resolution of your model or your null model or your way to produce the null model is going to vary.

So let us start with the simple example. Let us say that I want to just keep the number of nodes and edges constant. What would you do? I would just generate hurdle model by saying n is this and p is e/n choose 2. And I would be able to create a graph with roughly the same number of nodes and edges. It is varying by a little as we will see in the lab later today. It will vary by a little, but you will find that it is approximately represents the original network.

But then what happens, you can only say that with respect to numbers now. For this size network, I expect so many triangles to be there. So many feed forward loops to be there, but then that is the way poor null model to compare. It is like comparing your paracetamol against you know, may be taking another bad or a negative thing. But if you want to establish a more solid control you would say that I want the degree distribution also to be constant.

(Refer Slide Time: 07:23)



So how would you rewire a network or how would you create alternative networks. By the way these null models are important in every scenario. If you want to define how would a community is? or in fact, there is a definition of modularity that is based on these kinds of null models. So modularity can actually be defined as the number of intragroup edges - rest of the edges. Fraction of intragroup edges - the fraction of other edges.

So there are definitions of modularity that are based on this. So let us say I want to keep degree distribution unaltered. How would I go about creating a null model? How do I create a new network or I do not want to create 1 network I may want to generate a few 1000 networks? It is very similar to the idea we saw yesterday. We said that this is going to be the distribution of some LWS or rather LER and my LWS comes here.

So, this is this P will be much < say 10 to - 5 or something like that so it is statistically significant which means that I can say that LWS if I observe this. This actually we usually do not observe it for characteristic path length. So characteristic path length is whereas you know clustering coefficient you will find that this actually happens. So you have to, but asking for degree distribution to be constant would not you would not be being going to specific.

So how would you create a new network that has a same degree distribution as the original network? You can actually resolve to some sort of rewiring like your Watts Strogatz. So Watts Strogatz was removing a particular node here and connecting so in fact that is not the idea is rewiring. So let us say this is your network. You have AB and CD. One way to rewire this would be you now remove AB, CD and make it AC, BD or you make it AD, BC.

This will obviously preserve the number of edges or will it? Well the degree distribution. Has only 1 differential for degree distribution so it is N(k) versus K. will this graph change could be like this, it could be whichever way. See what was it is simple. So what happens here after rewiring it does not change, but is there a possible problem. I am always go to pick a pair and rewire them and I can do multiple rewirings that is not a problem.

So I will do enough rewiring so that I lose the memory of my original network. If I do 1 rewiring I cannot call it a random network. It is still a very close cousin of the original network we started off with, but if you do multiple rows of rewiring, you can get a substantially different network. So you could call it almost random variant of the original network and you repeat this process 1000 times you will get a you can compute a distribution of this sort.

There is one small problem here. What is you already had an edge like this. You cannot add a second edge between A and C, it does not work that way. So this could be a problem. So we you may have to account for that. You may want to do a few more simulations; few more iterations wherein you ignore if they already have an edge do not it you try to rewire some other pair and so on. There is another way to do it. So here you have a problem.

Here the thing is you cannot add a second edge. Here I cannot rewire AC and BD now so let us say it is already connected like this you have a problem. Now I cannot just say remove this and this and instead connect this. This means that I am going to add a second edge between C and D which is not even legal so there is a problem there so you have to basically find another pair of another set of ABCD and repeat this exercise, but in a sparks network it is not going to too much of a problem.

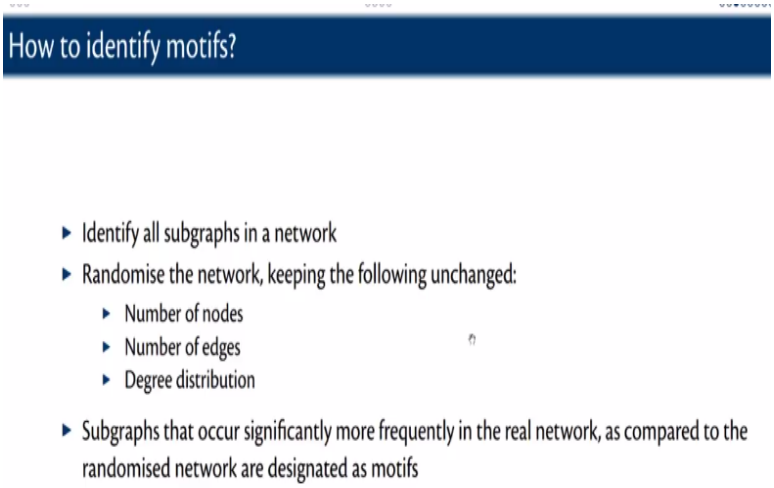
One other way to go about is think of these as threads and beads. Now you just cut the edges here leaving behind stubbs and now you tie any 2 threads together. This is what is the number of stubbs tell you degree. So you are not changing it. There is 1 problem here can you think of it. You do not want to left up left behind with this. Suppose you tied everything up when you are left with having to tie a node to itself.

But you can either permit that and have the appropriate null model or you know basically reject it and try to come up with a different solution. So, all these are basically algorithms. So while you have not done n number of rewirings try 1 rewiring. If that rewiring is acceptable change the graph to the newly obtained graph repeat. Leave the algorithm and you can go about it either for this kind of a rewiring or making these stubbs and rewiring.

(Refer Slide Time: 14:12)

So last time we were looking at how we need to randomize the network? What are the methods that we would use to randomize the network and what are the things you would keep constant?

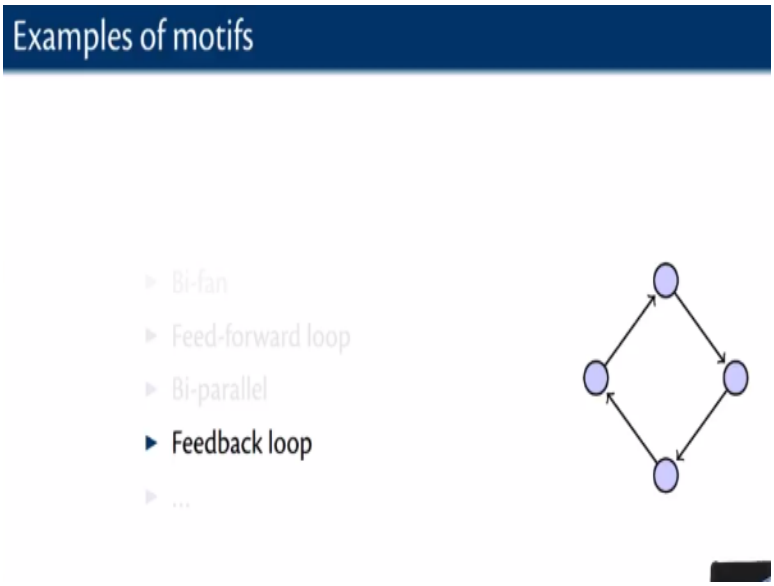
(Refer Slide Time: 14:22)



To identify motifs one needs to identify all the subgraphs in the network and you need to identify the statistically significantly over-represented subgraphs. For this you need multiple realizations of the original network, multiple random what one calls bootstrap, variations of the original network. So how would you randomize the network you want to keep the number of nodes same, the number of edges same, as well as the degree distribution.

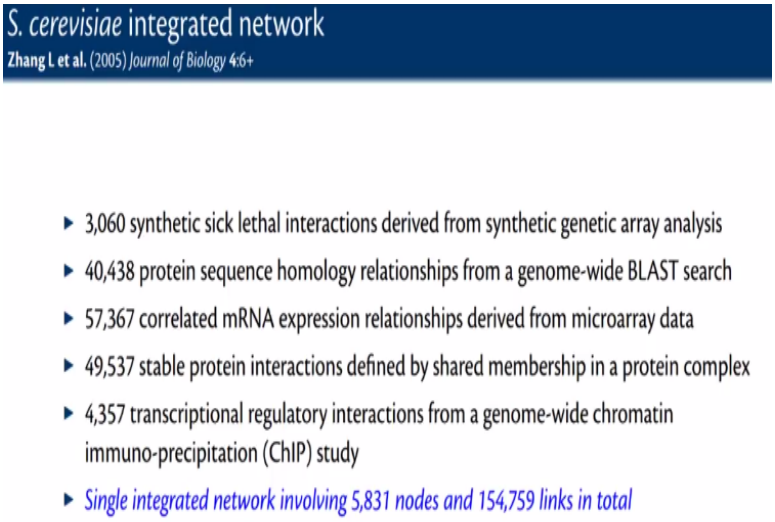
For this you discussed a couple of strategies to randomize the network and following this procedure the subgraphs that other significantly more frequently in the real network compared to random will be designated as motifs.

(Refer Slide Time: 15:05)



So what kind of motifs 1 might observe. So you might observe something that is known as a bi-fan that is seen here or a feed-forward loop, a bi-parallel motif or a feedback loop and you can have many more different kinds of motifs. One thing you can image is that number of motifs exponentially increases with the size of the motifs and of course the size of the network the number of motifs that you can potentially find.

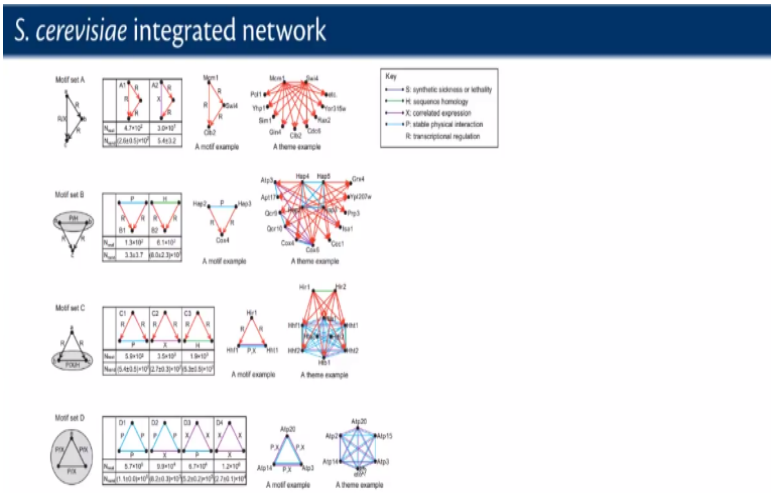
(Refer Slide Time: 15:39)



So Zhang and coworkers studied the east integrated network which contained several different interactions. So first was synthetic sick lethal interactions derived from a synthetic genetic array so essentially this contains interactions between genes that when together deleted cause the organism to die or at least because it falls sick meaning grows low or have some visible phenotypic damage.

And then 40,000 protein sequence homology relationships, correlated mRNA expressions, protein interactions based on protein complexations, and a ChiP study to get some more regulatory interactions and so on. Finally, giving you a huge network at about 6000 nodes and 150,000 links. In this network what was the motifs that could be identified.

(Refer Slide Time: 16:32)



(a) A motif corresponding to the 'feed-forward' theme; (b) motifs corresponding to the 'co-pointing' theme; (c) motifs corresponding to the 'regulonic complex' theme; (d) motifs corresponding to the 'protein complex' theme;

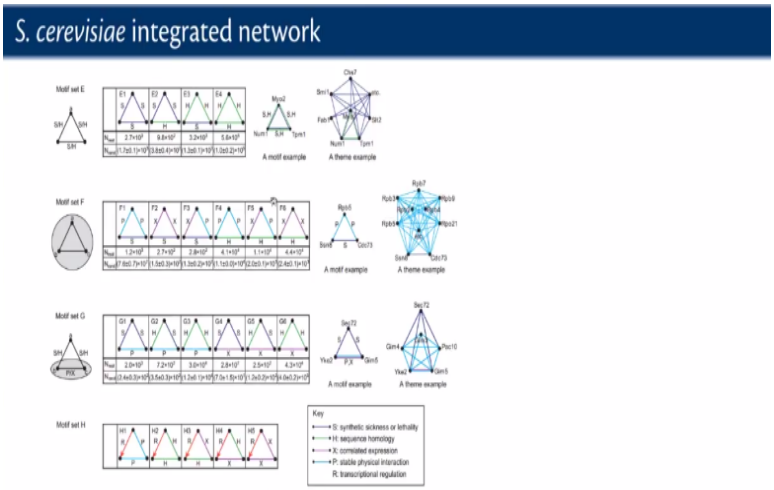
So this was 1 bunch of motifs. So if you can see this is a feed forward loop and what you find is they also have different edges here. Tag them as SHXPR and so on. So what you see here is the number of real networks were 4.7 * 10 to the power 2 whereas the number of random was about 260. So 260 + or - 50. So it turns out that this is significantly statistically significant much higher than the random. So the first motifs correspond to the feed-forward theme.

There are other motifs corresponding to the co-pointing theme, regulonic complex and so on and so forth. So all these you can see have different significances, different patterns and if you dig deeper they also have different colours which correspond to different types of interactions. So

there is a nice granular description of the motifs in this case. It is a very nice study that was published in the journal of biology in 2005.

Any functionality for the individual Motifs? So all these motifs I think they have gone to show that they have some interesting functions and so on. So many of these motifs have so if you see all these have some ATP banding proteins and things like that. So they are able to capture the biology as well.

(Refer Slide Time: 18:03)

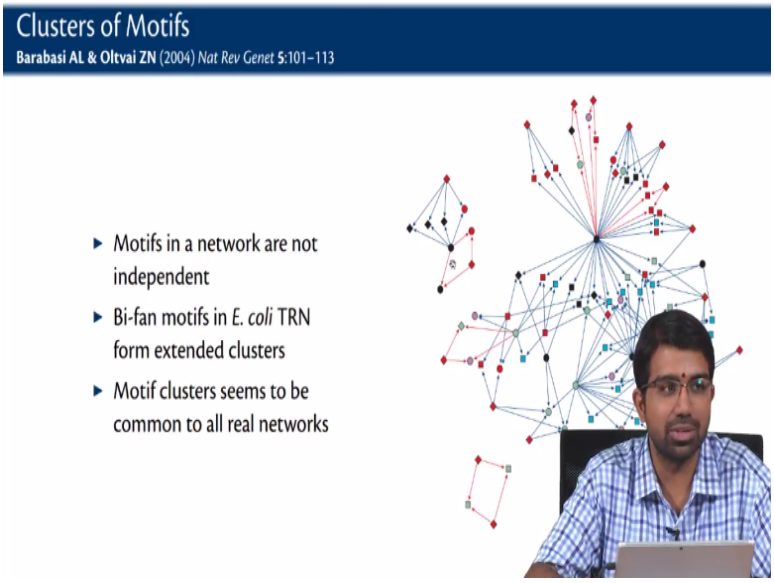


(e) motifs corresponding to the theme of neighborhood clustering of the integrated SSL/homology network; (f) motifs corresponding to the 'compensatory complex members' theme; (g) motifs corresponding to the 'compensatory protein and complex/process' theme; (h) other unclassified motifs

"Professor - student conversation starts" (()) (18:02) These are modules. Can I say that functionality of a Motifs it governs the functionality of the network? If number of motifs are large in a network? Yes, it is one of the you know, driving factors. The presence of a larger number of certain types of motifs may give you more robustness to the network and things like that certainly. **"Professor - student conversation ends"**

Here you have more examples where they report the number of real motifs as well as the number found in random motifs and you know the compensate recomplexes, unclassified motifs, neighbourhood clusters so on and so forth. They basically identify different types of motifs in these networks.

(Refer Slide Time: 18:47)

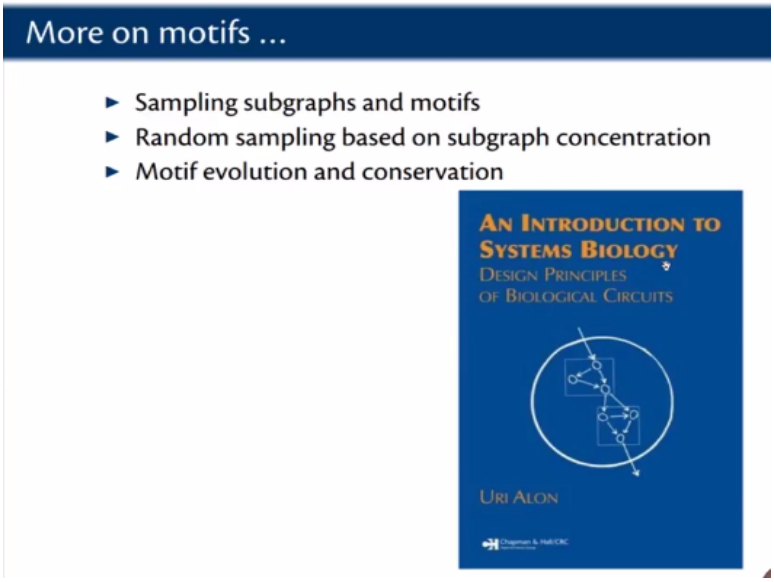


The 1 important thing to remember is that motifs are not independent as you were just mentioning motifs form clusters themselves. So they will find a bi-fan motifs and E. coli transcription regulatory network form extended clusters. So many of these are bi-fans and they form extended clusters to recap a bi-fan looks like this. This is the bifan. And motif cluster seemed to be common to all real networks.

So in some sense we have understood very well how to identify motifs. The first thing it is not simple, but when you get a practice you will see what are the challenges? So the first thing is you have a network. You identify all subgraphs in the network and then you count them and then you generate randomized realizations of the same network count again and then plot distributions and find out.

Or just do counts and averages and say you can even do a parametric test if you want to say that your number of motifs in the real network is statistically significant, but what is the challenge here. Computational. How do you enumerate all subgraphs in a network? It is going to be massive, millions, billons I mean uncountable subgraphs.

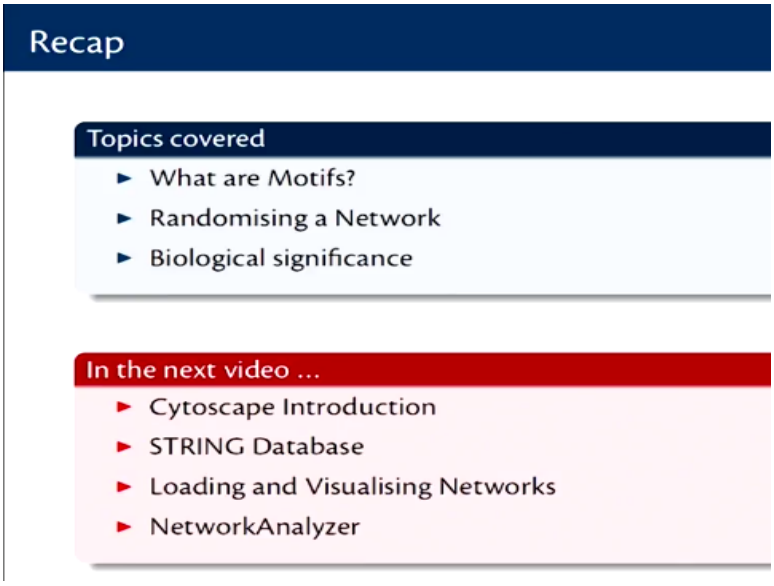
(Refer Slide Time: 20:07)



So then you have to resort to some interesting things like sampling subgraphs. So you try to sample parts of the network to find subgraphs and motifs and you actually have intelligent methods that do random sampling based on subgraph concentration and one also is very interested in studying how motifs evolve.

And other conserved motifs across species, across kingdoms of life across organism, across different types of networks you could have a co-expression network versus a metabolic network versus a genetic regulatory network and so on. The Uri Alon has this very classic text books on systems biology which covers a lot of information on motifs and networks.

(Refer Slide Time: 20:48)



So I hope you got a good glimpse of what motifs are? Some concepts related to randomization, statistical hypothesis testings in terms of identifying motifs in a network as well as how these motifs are embedded in biological networks. I really invite you to go and read a little more about motifs in the book by Uri Alon I just showed and in the next video, we will move towards the Lab session wherein I will introduce you to cytoscape, talk to you about some concepts of underlying the string data base as well as how do you load and visualize networks and use this very interesting cytoscape tool will called network analyzer.